

MKL-RT: Multiple Kernel Learning for Ratio-trace Problems via Convex Optimization

Raviteja Vemulapalli, Vinay Praneeth Boda, Rama Chellappa

University of Maryland, College Park

Abstract. In the recent past, automatic selection or combination of kernels (or features) based on multiple kernel learning (MKL) approaches has been receiving significant attention from various research communities. Though MKL has been extensively studied in the context of support vector machines (SVM), it is relatively less explored for ratio-trace problems. In this paper, we show that MKL can be formulated as a convex optimization problem for a general class of ratio-trace problems that encompasses many popular algorithms used in various computer vision applications. We also provide an optimization procedure that is guaranteed to converge to the global optimum of the proposed optimization problem. We experimentally demonstrate that the proposed MKL approach, which we refer to as MKL-RT, can be successfully used to select features for discriminative dimensionality reduction and cross-modal retrieval. We also show that the proposed convex MKL-RT approach performs better than the recently proposed non-convex MKL-DR approach.

1 Introduction

In many computer vision applications, we are often interested in transforming our initial feature representation x to a new representation x' such that x' suits better for the application under consideration. For example, if we are interested in classification, we may want to transform x such that samples from the same class are close to each other and samples from different classes are far away from each other. If we are interested in retrieving images using text query, we may want to transform our image representation such that it is highly correlated with the corresponding text representation.

If we plan to use a linear transformation, then we are interested in learning a transformation matrix W , such that $x' = W^\top x$ has certain desired properties depending on the application of interest. Though various different algorithms (aiming at different applications) have been proposed in the past for learning the transformation matrix W , many of them end up solving a ratio-trace problem [40,37] (equation (2)), whose optimal solution can be obtained using generalized Eigenvalue decomposition (GEVD) [10].

Algorithms based on ratio-trace problems have been extensively used in various computer vision applications [40,8,3,32,5,17,7,19,30]. Some of the popular algorithms formulated as a ratio-trace problem (equation (2)) are linear discriminant analysis (LDA) [10], semi-supervised discriminant analysis (SDA) [5], side-information based LDA (SILDA) [17], local discriminant embedding (LDE) [7], marginal Fisher analysis (MFA) [40], locality preserving projections (LPP) [15], neighborhood preserving embedding (NPE) [14], canonical correlation analysis (CCA) [13], and orthonormal PLS-SB [31].

All the above mentioned linear algorithms suffer from two main disadvantages:

(i) They require input data to be represented in the form of feature vectors x in an Euclidean space. This may not be possible in applications where the data of interest is represented using bag-of-features [22], matrices [34] or manifold features [36]. In some applications, we may only have similarities or distances between the features instead of explicit representations. (ii) Linear transformations may be too simple to be effective in some applications as they can not handle non-linearity present in the data. Both of these issues can be handled by using kernels. Kernelized versions of these linear algorithms also end up solving a ratio-trace problem (equation (3)).

Though kernel-based methods have been successfully used in many computer vision applications, the kernel function and the associated feature space are central choices that are generally made by the user. Recently, automatic selection or combination of kernels (or features) based on MKL approaches has been shown to produce state-of-the-art results in various applications [35,11,42]. Multiple kernel learning was initially proposed [21] for SVM and has since received significant attention [33,29,20]. An excellent overview of various MKL algorithms can be found in [12].

Though MKL has been extensively studied in the context of SVM, it is relatively less explored for ratio-trace problems. Motivated by MKL-SVM, Kim *et. al.* [18] and Ye *et. al.* [41] extended the MKL approach to LDA (which is a specific instance of ratio-trace problem (2)) formulating it as a convex optimization problem. Arguing for non-sparse MKL, Yan *et. al.* [39] proposed a non-sparse version of MKL-LDA, which imposes a general ℓ_p norm regularization on the kernel weights.

Motivated by MKL-LDA, Lin *et. al.* [23] extended the MKL approach to graph embedding framework [40] which covers a large number of dimensionality reduction algorithms. The MKL-DR framework of [23] is based on trace-ratio formulation which is different from the ratio-trace formulation used in this paper. We refer the readers to [37] for a discussion on the differences between trace-ratio and ratio-trace formulations. The trace-ratio based MKL-DR formulation of [23] results in a non-convex optimization problem. In [23], the authors used an iterative optimization procedure that has no convergence guarantees.

In this paper, we show that similar to MKL-LDA [41], kernel learning can be formulated as a convex optimization problem for a large class of ratio-trace problems (equations (2) and (3)) that includes popular algorithms like LDA, SDA, SILDA, LDE, NPE, MFA, LPP, CCA and orthonormal PLS-SB. We also provide an optimization procedure that is guaranteed to converge to the global optimum of the proposed convex optimization problem.

In practice, MKL is typically used in two different ways: (i) Various kernels are defined for the same feature representation, for example Gaussian kernels with different values of σ , and an optimal kernel is learned as a combination of these kernels [21,41], (ii) Multiple feature descriptors are used to represent objects of interest and a similarity kernel is generated from each feature [11,23]. In this case, kernel learning effectively solves the feature selection problem and the MKL coefficients can be interpreted as weights given to the corresponding features. One can also use a mixed strategy [42] of using multiple features and defining multiple kernels for each feature. In this paper, we use MKL for feature selection in the context of ratio-trace problems.

Contributions: 1) We show that MKL can be formulated as a convex optimization problem for a large class of ratio-trace problems. The proposed MKL-RT formulation is applicable to various popular algorithms like LDA, SDA, SILDA, LDE, NPE, MFA, LPP, CCA and orthonormal PLS-SB. 2) We provide an optimization procedure that is guaranteed to converge to the global optimum of the proposed optimization problem. 3) We experimentally show that the proposed MKL-RT approach can be successfully used to select features for discriminative dimensionality reduction and cross-modal retrieval. 4) We show that the proposed ratio-trace based convex MKL-RT approach performs better than the trace-ratio based non-convex MKL-DR approach of [23].

Organization: Section 2 presents the general class of ratio-trace problems for which MKL can be formulated as a convex optimization problem. Section 3 discusses some specific instances of the ratio-trace problem which will be used in our experimental evaluation. Section 4 presents the proposed convex MKL-RT formulation. We present our experimental results in section 5 and conclude the paper in section 6.

Notations: We use $\mathbb{1}$ to denote the indicator function. The transpose of a matrix A is denoted by A^\top . We use I to denote an identity matrix of appropriate size. We use $|\cdot|$ to denote the absolute value and \emptyset to denote the empty set.

2 Ratio-trace Problem

Definition 1: For any two $d \times d$ symmetric positive semi-definite matrices S_1 and S_2 , the ratio-trace problem is defined as

$$\underset{W}{\text{maximize}} \quad \text{trace} \left[(W^\top S_1 W)^{-1} (W^\top S_2 W) \right]. \quad (1)$$

In this paper, we focus on the class of algorithms that learn the data transformation matrix W by solving the following ratio-trace problem:

$$\underset{W}{\text{maximize}} \quad \text{trace} \left[(W^\top ((1 - \sigma)XLX^\top + \sigma I) W)^{-1} (W^\top XL'X^\top W) \right], \quad (2)$$

where $X = [x_1, \dots, x_N] \in \mathcal{R}^{d \times N}$ is the data matrix (assumed to be centered), $\sigma \in (0, 1)$ is a regularization parameter used to prevent overfitting, and $L (\neq 0)$, L' are (algorithm-dependent) $N \times N$ symmetric positive semi-definite matrices.

Some of the popular algorithms that fall into this class are LDA, SDA, SILDA, LDE, MFA, LPP, NPE, CCA and orthonormal PLS-SB. All these linear algorithms can be made non-linear by using kernels. For a given kernel function \mathcal{K} , the kernelized versions of these algorithms solve the following ratio-trace problem:

$$\underset{\Gamma}{\text{maximize}} \quad \text{trace} \left[(\Gamma^\top ((1 - \sigma)K L K + \sigma K) \Gamma)^{-1} (\Gamma^\top K L' K \Gamma) \right], \quad (3)$$

where K is the $N \times N$ kernel matrix with $K_{ij} = \mathcal{K}(x_i, x_j)$. The optimal solution to (3) is given by the generalized Eigenvectors [10] corresponding to the non-zero generalized Eigenvalues of the matrix pair $(K L' K, (1 - \sigma)K L K + \sigma K)$:

$$K L' K \gamma = \lambda ((1 - \sigma)K L K + \sigma K) \gamma. \quad (4)$$

Once Γ is obtained, the new (non-linearly transformed) representation for a data sample $x \in \mathcal{R}^d$ can be computed using $x' = \Gamma^\top [\mathcal{K}(x_1, x), \dots, \mathcal{K}(x_N, x)]^\top$.

3 Instances of the Ratio-trace Problem

As mentioned earlier, various algorithms [40,5,17,7,15,14,13,31] used in many computer vision applications are formulated as a ratio-trace problem. In this section, we briefly discuss three specific instances of the ratio-trace problem which will be used in the experimental evaluation of the proposed MKL-RT approach in section 5.

3.1 KFDA - Kernel Fisher Discriminant Analysis

KFDA [2] is a popularly-used non-linear discriminative dimensionality reduction algorithm. Let $\{(x_i, y_i), i = 1, \dots, N\}$ be the set of N labeled training samples, where $y_i \in \{1, \dots, P\}$ is the class label of feature $x_i \in \mathcal{R}^d$. Let $y = [y_1, \dots, y_N]$, $\mathbf{1}_p = \mathbb{1}[y = p]$ be the membership vector corresponding to class p , and N_p be the number of labeled samples in class p . Let \mathcal{K} be a kernel function defined on features x_i and K^x be the corresponding kernel matrix. Then, KFDA solves the ratio-trace problem (3) with

$$K = K^x, \quad L' = \sum_{p=1}^P \frac{1}{N_p} \mathbf{1}_p \mathbf{1}_p^\top, \quad \text{and} \quad L = I - L' \text{ or } L = I. \quad (5)$$

The lower dimensional representation for a data sample $x \in \mathcal{R}^d$ can be computed using $x' = \Gamma^\top [\mathcal{K}(x_1, x), \dots, \mathcal{K}(x_N, x)]^\top$.

3.2 KCCA - Kernel Canonical Correlation Analysis

KCCA [13] is a popular approach used in cross-modal retrieval applications. KCCA maps the data (non-linearly) from two different modalities to a common lower dimensional latent/concept space where the two modalities are highly correlated. Let $\{(x_i, z_i), i = 1, \dots, N\}$ be N training data pairs where $x_i \in \mathcal{R}^{d_1}$ and $z_i \in \mathcal{R}^{d_2}$ are samples from the first and second modalities respectively. Let \mathcal{K}^x and \mathcal{K}^z be kernel functions defined on features x_i and z_i respectively. Let K^x and K^z be the corresponding kernel matrices. Then, KCCA solves the ratio-trace problem (3) with

$$K = K^x, \quad L' = K^z ((1 - \sigma)K^z + \sigma I)^{-1} \quad \text{and} \quad L = I. \quad (6)$$

The latent space representations for samples $x \in \mathcal{R}^{d_1}$ and $z \in \mathcal{R}^{d_2}$ from first and second modalities respectively are given by $x' = \Gamma^\top [\mathcal{K}^x(x_1, x), \dots, \mathcal{K}^x(x_N, x)]^\top$ and $z' = \Xi^\top [\mathcal{K}^z(z_1, z), \dots, \mathcal{K}^z(z_N, z)]^\top$, where

$$\Xi = ((1 - \sigma)K^z + \sigma I)^{-1} K^x \Gamma \Lambda^{-\frac{1}{2}}. \quad (7)$$

Here, Λ is the diagonal matrix of non-zero generalized Eigenvalues given by (4).

3.3 LKCCA - Labeled Kernel Canonical Correlation Analysis

KCCA requires paired training data samples $\{(x_i, z_i)\}$ from two modalities to learn the transformations from the initial feature spaces to the common latent space. Suppose,

instead of pairings, we are provided with class labels for the training data in the two modalities. We cannot directly use KCCA in this case. One simple way to handle this situation is to generate data pairs using the class labels and then use KCCA with the generated pairs. We refer to this extension of KCCA as labeled KCCA in this paper.

Let $\{(x_i, y_i), i = 1, \dots, N^x\}$ be N^x labeled training samples from first modality with $y_i \in \{1, \dots, P\}$ being the class label of feature x_i . Let $\{(z_j, w_j), j = 1, \dots, N^z\}$ be N^z labeled training samples from second modality with $w_j \in \{1, \dots, P\}$ being the class label of feature z_j . Let N_p^x and N_p^z denote the number of training samples from class p in first and second modalities respectively. Let $y = [y_1, \dots, y_{N^x}]$ and $\mathbf{1}_p^x = \mathbb{1}[y = p]$. Let $w = [w_1, \dots, w_{N^z}]$ and $\mathbf{1}_p^z = \mathbb{1}[w = p]$.

In LKCCA, we form a training pair between samples x_i and z_j if $y_i = w_j$. A straightforward way to implement this is to replicate each data sample as many times as the number of samples from the same class in the other modality. But, this would unnecessarily increase the size of kernel matrices K^x and K^z . Instead, LKCCA can be efficiently implemented without actually replicating the samples. This efficient implementation of LKCCA solves the ratio-trace problem (3) with

$$K = K^x, \quad L' = EK^z(\sigma I + (1 - \sigma)D^z K^z)^{-1}E^\top \quad \text{and} \quad L = D^x, \quad (8)$$

where D^x is a $N^x \times N^x$ diagonal matrix with $D_{i,i}^x = N_{y_i}^z$, D^z is a $N^z \times N^z$ diagonal matrix with $D_{j,j}^z = N_{w_j}^x$, and $E = \sum_{p=1}^P \mathbf{1}_p^x \mathbf{1}_p^{z\top}$.

The latent space representations for samples $x \in \mathcal{R}^{d_1}$ and $z \in \mathcal{R}^{d_2}$ from first and second modalities respectively are given by $x' = \Gamma^\top [\mathcal{K}^x(x_1, x), \dots, \mathcal{K}^x(x_{N^x}, x)]^\top$ and $z' = \Xi^\top [\mathcal{K}^z(z_1, z), \dots, \mathcal{K}^z(z_{N^z}, z)]^\top$, where

$$\Xi = ((1 - \sigma)D^z K^z + \sigma I)^{-1}E^\top K^x \Gamma \Lambda^{-\frac{1}{2}}. \quad (9)$$

Here, Λ is the diagonal matrix of non-zero generalized Eigenvalues given by (4). We refer the readers to supplementary material for further details about LKCCA.

4 MKL-RT: MKL for Ratio-trace Problem

In the MKL framework, the kernel function \mathcal{K} is parametrized as a linear combination of pre-defined base kernel functions $\mathcal{K}^1, \dots, \mathcal{K}^M$:

$$\mathcal{K} = \sum_{m=1}^M \mu_m \mathcal{K}^m, \quad \mu_m \geq 0, \quad \sum_{m=1}^M \mu_m = 1, \quad (10)$$

and the weights $\mu = [\mu_1, \dots, \mu_M]$ are learned from the data. Under this framework, MKL for ratio-trace problem can be formulated as the following optimization problem:

$$\begin{aligned} & \underset{\Gamma, K, \mu}{\text{maximize}} \quad \text{trace} \left[(\Gamma^\top ((1 - \sigma)K L K + \sigma K) \Gamma)^{-1} (\Gamma^\top K L' K \Gamma) \right] \\ & K = \sum_{m=1}^M \mu_m K^m, \quad \sum_{m=1}^M \mu_m = 1, \quad \mu_m \geq 0 \quad \forall m, \end{aligned} \quad (11)$$

where K^m is the kernel matrix corresponding to the base kernel function \mathcal{K}^m .

The optimization problem (11) is a non-convex optimization problem. Nevertheless, the optimal μ^* for (11) can be obtained by solving a different convex optimization problem as stated in the following theorem.

Theorem 1: Let L and L' be two symmetric positive semi-definite matrices with ranks l and l' respectively. Let $\{(\alpha_i, u_i)\}_{i=1}^l$ and $\{(\beta_i, v_i)\}_{i=1}^{l'}$ be the non-zero Eigenvalue-Eigenvector pairs of L and L' respectively. Let $G = [\sqrt{\alpha_1}u_1, \dots, \sqrt{\alpha_l}u_l]$ and $h_i = \sqrt{\beta_i}v_i$ for $i = 1, 2, \dots, l'$. For $m = 1, \dots, M$, let

$$S_m(\eta) = \frac{1}{\sigma} \sum_{i=1}^{l'} \left(\frac{\eta_i^\top \eta_i}{4(1-\sigma)} + \frac{\eta_i^\top G^\top K^m G \eta_i}{4\sigma} - \eta_i^\top G^\top K^m h_i \right) + \text{trace}(K^m L') \quad (12)$$

be M functions defined $\forall \eta = [\eta_1, \dots, \eta_{l'}] \in \mathcal{R}^{l \times l'}$. Let μ^* be optimal for the following convex optimization problem:

$$\begin{aligned} & \underset{\zeta, \mu}{\text{maximize}} \quad \zeta \\ & \text{subject to} \quad \sum_{m=1}^M \mu_m S_m(\eta) \geq \zeta, \quad \forall \eta \in \mathcal{R}^{l \times l'}, \\ & \quad \sum_{m=1}^M \mu_m = 1, \quad \mu_m \geq 0 \quad \forall m. \end{aligned} \quad (13)$$

Then μ^* is optimal for the optimization problem (11).

Proof: Please refer to the supplementary material for the proof.

Note that the optimization problem (13) is a semi-infinite linear program (SILP). Following [33,41], we use an iterative approach to solve (13). In each iteration, the optimal μ and ζ are computed for a restricted subset of constraints C in (13). Constraints that are not satisfied by current μ and ζ are added successively to the restricted problem until all the constraints are satisfied. For faster convergence, in each iteration, we add the constraint that maximizes the violation for current μ and ζ . To find the maximum violating constraint, we solve

$$\eta^* = \underset{\eta}{\text{argmin}} \left(\sum_{m=1}^M \mu_m S_m(\eta) - \zeta \right). \quad (14)$$

Using the definition of $S_m(\eta)$ from equation (12), it can be easily verified that the optimum $\eta^* = [\eta_1^*, \eta_2^*, \dots, \eta_{l'}^*]$ for (14) can be obtained by individually solving for each η_i^* using

$$\eta_i^* = \underset{\eta_i}{\text{argmin}} \left(\frac{\eta_i^\top \eta_i}{4\sigma(1-\sigma)} + \frac{\eta_i^\top G^\top K G \eta_i}{4\sigma^2} - \frac{\eta_i^\top G^\top K h_i}{\sigma} \right), \quad (15)$$

where $K = \sum_{m=1}^M \mu_m K^m$. Note that the optimization problem (15) is an unconstrained quadratic program whose solution can be obtained by solving the following

Table 1: Algorithm for solving SILP (13).

Input: $L', G, \{h_i\}_{i=1}^{l'}, \{K^m\}_{m=1}^M, T, \sigma, \epsilon.$
Initialization: $\mu_m = \frac{1}{M} \forall m, \zeta = +\infty, t = 1, C = \emptyset.$
while $t \leq T$ $K = \sum_{m=1}^M \mu_m K^m$ for $i = 1, \dots, l'$ Compute η_i^* by solving $\left(\frac{I}{2(1-\sigma)} + \frac{G^\top K G}{2\sigma}\right) \eta_i^* = G^\top K h_i.$ end if $\left 1 - \frac{\sum_{m=1}^M \mu_m S_m(\eta^*)}{\zeta}\right < \epsilon$ break; else Add η^* to the constraint set C . Update μ and ζ by solving restricted version of (13) using only $\eta \in C$. end $t = t + 1;$ end
Output: μ^* and $\zeta^*.$

system of linear equations:

$$\left(\frac{I}{2(1-\sigma)} + \frac{G^\top K G}{2\sigma}\right) \eta_i^* = G^\top K h_i. \quad (16)$$

Hence, in each iteration we solve l' linear systems to find the maximum violating constraint and one linear program to update μ and ζ . Following [33,41], we use the following stopping criterion:

$$\left|1 - \frac{\sum_{m=1}^M \mu_m S_m(\eta^*)}{\zeta}\right| < \epsilon. \quad (17)$$

Table 1 summarizes the algorithm used for solving (13). This iterative algorithm is referred to as column generation technique and is guaranteed to converge [33,41].

Once the optimal μ^* is obtained, we can solve the ratio-trace problem (3) using GEVD with $K^* = \sum_{m=1}^M \mu_m^* K^m$ to get the optimal Γ^* . Once μ^* and Γ^* are known, the new non-linearly transformed representation for a data sample $x \in \mathcal{R}^d$ can be computed using $x' = \Gamma^{*\top} [\mathcal{K}^*(x_1, x), \dots, \mathcal{K}^*(x_N, x)]^\top$, where $\mathcal{K}^*(x_i, x) = \sum_{m=1}^M \mu_m^* \mathcal{K}^m(x_i, x)$. Table 2 summarizes the proposed MKL-RT algorithm.

5 Experimental Evaluation

We evaluated the proposed convex MKL-RT approach using three different instances of the ratio-trace problem: KFDA, KCCA and LKCCA (explained in section 3), covering two different applications: discriminative dimensionality reduction (for classification) and cross-modal retrieval. We used Caltech101 [9] and Oxford flowers17 [25] datasets for discriminative dimensionality reduction experiments, and Wikipedia articles [30] and PascalVOC 2007 [16] datasets for cross-modal retrieval experiments.

Table 2: MKL-RT algorithm.

Input: $L', L, \{\mathcal{K}^m\}_{m=1}^M, T, \sigma, \epsilon$.
Compute $G = [\sqrt{\alpha_1}u_1, \dots, \sqrt{\alpha_{l'}}u_{l'}]$, and $h_i = \sqrt{\beta_i}v_i$ for $i = 1, \dots, l'$, where $\{(\alpha_i, u_i)\}_{i=1}^{l'}$ and $\{(\beta_i, v_i)\}_{i=1}^{l'}$ are the non-zero Eigenvalue-Eigenvector pairs of L and L' respectively.
Solve the SILP (13) to obtain μ^* using the algorithm summarized in table 1.
Solve the ratio-trace problem (3) using GEVD with $K^* = \sum_{m=1}^M \mu_m^* K^m$ to get optimal Γ^* .
The new non-linearly transformed representation for a data sample $x \in \mathcal{R}^d$ can be computed using $x' = \Gamma^{*\top} [\mathcal{K}^*(x_1, x), \dots, \mathcal{K}^*(x_N, x)]^\top$, where $\mathcal{K}^*(x_i, x) = \sum_{m=1}^M \mu_m^* \mathcal{K}^m(x_i, x)$.

5.1 Comparative Methods

In all the experiments, we compare the proposed MKL-RT approach with the following methods:

- **AK-RT (Average kernel):** We solve the ratio-trace problem (3) using the arithmetic mean kernel \mathcal{K}_A defined as $\mathcal{K}_A = \frac{1}{M} \sum_{m=1}^M \mathcal{K}^m$.
- **PK-RT (Product kernel):** We solve the ratio-trace problem (3) using the geometric mean kernel \mathcal{K}_P defined as $\mathcal{K}_P = (\prod_{m=1}^M \mathcal{K}^m)^{1/M}$.
- **BIK-RT (Best individual kernel):** We solve the ratio-trace problem (3) using the best kernel among $\{\mathcal{K}^m\}_{m=1}^M$.
- **Non-convex MKL-DR¹:** We use the trace-ratio based non-convex MKL approach proposed in [23].

In the case of Caltech101 and Oxford flowers17 datasets, we use SVM and nearest neighbor (NN) rule for classification after discriminative dimensionality reduction. Hence, for these datasets, we also compare the proposed MKL-RT approach with the following SVM and NN based approaches:

- **Kernel SVM approaches:** Average kernel SVM (AK-SVM), product kernel SVM (PK-SVM), best individual kernel SVM (BIK-SVM) and MKL-SVM.
- **Kernel NN approaches (without dimensionality reduction):** Average kernel NN (AK-NN), product kernel NN (PK-NN) and best individual kernel NN (BIK-NN). We computed the distances from kernels using

$$d^2(x_i, x_j) = \mathcal{K}(x_i, x_i) + \mathcal{K}(x_j, x_j) - \mathcal{K}(x_i, x_j) - \mathcal{K}(x_j, x_i). \quad (18)$$

5.2 KFDA for Discriminative Dimensionality Reduction

In these experiments, we first performed dimensionality reduction using KFDA and then used SVM and NN rule for classification in the lower dimensional space. We used two different datasets, namely Caltech101 [9] and Oxford flowers17 [25]. The number of KFDA dimensions was chosen to be $P - 1$, where P is the number of classes.

¹ We used the code obtained from the authors of [23] through personal correspondence.

Caltech101 [9] is a multiclass object recognition dataset with 101 object categories and 1 background category. The authors of [11] precomputed 39 different kernels for this dataset using various image features and the kernel matrices are available online². We used these 39 kernels in our experiments and followed the experimental setup used in [11]. For brevity we omit the details of the features and refer to [11]. We report the results using all 102 classes of Caltech101. We repeated the experiment 5 times using different training and test splits and report the average results. The performance measure used is the mean prediction rate per class. We performed experiments using 5, 10, 15, 20, 25 and 30 training images per class and up to 50 test images per class. The regularization parameter σ was chosen based on cross-validation.

Table 3 shows the recognition rates of various approaches for this dataset. For AK-SVM, PK-SVM, BIK-SVM and MKL-SVM, we report the results from [11], which were obtained using the same kernel matrices and splits. We can make the following key observations from these results:

- Simple kernel NN approaches produce very poor results.
- Performing discriminative dimensionality reduction gives a huge improvement with NN classifier (around 25-30%) and a moderate improvement with SVM classifier (around 2%). This is expected since the dimensionality reduction step makes samples from same class to be close to each other and samples from different classes to be far away from each other.
- Among various KFDA approaches, the proposed convex MKL-RT approach works best with both NN and SVM classifiers.
- The non-convex MKL-DR approach of [23] performs poorly compared to AK-RT, PK-RT and MKL-RT approaches. The standard deviation of the MKL-DR approach is very high when the number of training samples is low (around 4% for 5 training samples per class and 2.5% for 10 training samples per class). This shows that the non-convex MKL-DR is overfitting the training data.
- The proposed MKL-RT approach performs better (around 2.5% on average) than MKL-SVM.

Table 4 shows the number of kernels selected by MKL-RT-KFDA for the 5 splits used in our experiments. A kernel \mathcal{K}^m is considered to be selected if its contribution is greater than 0.1%, i.e., its coefficient μ_m is greater than $1/1000$. We can see that the number of kernels selected by MKL-RT-KFDA (around 7-14) is much less than the total number of kernels, which is 39 in this case. This clearly shows that the proposed approach can be successfully used to select features for discriminative dimensionality reduction. In contrast, the non-convex MKL-DR approach of [23] ended up selecting all the 39 kernels (all the weights were greater than 0.001 after ℓ_1 normalization). The main reason for this could be the lack of sparsity-promoting ℓ_1 -norm constraint (on the weights) in the MKL-DR formulation. Figures 1 and 2 respectively show the kernel weights of MKL-RT-KFDA and MKL-DR-KFDA approaches for the fifth random split with 30 training samples per class.

² <http://files.is.tue.mpg.de/pghler/projects/iccv09/>

Table 3: Average recognition rates for Caltech101 dataset.

	Number of training images per class					
Method	5	10	15	20	25	30
AK-NN	27.3 \pm 0.5	32.6 \pm 0.6	36.1 \pm 0.9	38.0 \pm 1.1	39.9 \pm 1.0	42.4 \pm 1.5
PK-NN	27.5 \pm 0.7	32.8 \pm 1.0	36.4 \pm 1.1	38.3 \pm 1.2	40.4 \pm 1.1	42.9 \pm 1.4
BIK-NN	29.4 \pm 0.9	35.5 \pm 1.2	39.5 \pm 0.7	41.3 \pm 1.0	43.3 \pm 0.8	45.7 \pm 1.1
AK-RT-KFDA + NN	46.0 \pm 0.8	57.6 \pm 0.6	64.2 \pm 0.8	68.1 \pm 1.2	70.8 \pm 0.9	73.6 \pm 1.1
PK-RT-KFDA + NN	45.0 \pm 0.6	56.1 \pm 0.6	62.6 \pm 0.6	66.7 \pm 0.7	69.4 \pm 0.9	72.4 \pm 1.0
BIK-RT-KFDA + NN	46.1 \pm 0.6	54.8 \pm 0.1	59.7 \pm 0.5	63.0 \pm 0.8	65.4 \pm 0.7	67.8 \pm 0.7
MKL-RT-KFDA + NN	45.7 \pm 0.6	58.6 \pm 0.4	65.3 \pm 0.8	69.5 \pm 0.7	72.1 \pm 0.4	74.6 \pm 0.7
MKL-DR-KFDA + NN	40.2 \pm 4.0	53.6 \pm 2.5	61.7 \pm 1.4	65.5 \pm 0.8	68.6 \pm 0.9	72.1 \pm 1.2
AK-SVM	44.4 \pm 0.6	55.7 \pm 0.5	62.2 \pm 1.1	66.1 \pm 1.0	68.9 \pm 1.0	71.6 \pm 1.5
PK-SVM	43.6 \pm 0.7	54.7 \pm 0.5	61.3 \pm 0.9	65.4 \pm 0.8	68.3 \pm 0.7	71.3 \pm 1.4
BIK-SVM	46.1 \pm 0.9	55.6 \pm 0.5	61.0 \pm 0.2	64.3 \pm 0.9	66.9 \pm 0.8	69.4 \pm 0.4
MKL-SVM	42.1 \pm 1.2	55.1 \pm 0.7	62.3 \pm 0.8	67.1 \pm 0.9	70.5 \pm 0.8	73.7 \pm 0.7
AK-RT-KFDA + SVM	46.1 \pm 0.8	57.6 \pm 0.6	64.2 \pm 0.8	68.1 \pm 1.0	70.8 \pm 0.9	73.7 \pm 1.0
PK-RT-KFDA + SVM	45.0 \pm 0.6	56.2 \pm 0.6	62.6 \pm 0.7	66.7 \pm 0.8	69.4 \pm 0.9	72.5 \pm 1.1
BIK-RT-KFDA + SVM	46.1 \pm 0.6	54.8 \pm 0.2	59.8 \pm 0.4	63.0 \pm 0.8	65.5 \pm 0.6	67.8 \pm 0.7
MKL-RT-KFDA + SVM	46.3 \pm 0.9	58.9 \pm 0.4	65.5 \pm 0.7	69.7 \pm 1.0	72.2 \pm 0.4	74.7 \pm 0.7
MKL-DR-KFDA + SVM	40.2 \pm 4.0	53.6 \pm 2.5	61.7 \pm 1.4	65.5 \pm 0.8	68.6 \pm 0.9	72.1 \pm 1.3

Table 4: Number of kernels selected by MKL-RT-KFDA for Caltech101.

Split	Training images per class					
	5	10	15	20	25	30
1	9	9	10	11	11	11
2	7	9	8	9	13	12
3	9	10	10	11	12	12
4	10	11	12	12	13	14
5	8	10	10	11	11	11

Fig. 1: Weights of proposed convex MKL-RT-KFDA for Caltech101.

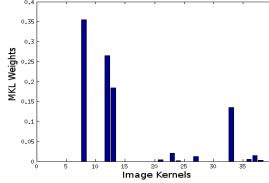
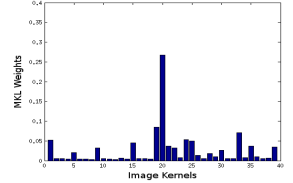


Fig. 2: Weights of non-convex MKL-DR-KFDA for Caltech101.



Oxford flowers17 [25] is a multiclass dataset consisting of 17 categories of flowers with 80 images per category. This dataset comes with 3 predefined splits into training (17×40 images), validation (17×20 images) and test (17×20 images) sets. Moreover, the authors of [25] precomputed 7 distance matrices using various features and the matrices are available online³. For brevity we omit the details of the features and refer to [25,26]. We used these distance matrices and followed the same procedure as in [11] to compute 7 different kernels: $\mathcal{K}^m(x_i, x_j) = \exp(-d_m(x_i, x_j)/\eta_m)$, where η_m is the mean of the pairwise distances d_m for the m^{th} feature. We performed experiments using the three predefined splits and report the average results. The regularization parameter σ was chosen based on cross-validation using the training and validation sets.

Table 5 shows the recognition rates of various approaches for this dataset. For the SVM-based approaches, we report the results from [11], which were obtained using the same kernel matrices and splits. We can see that all the observations (except the high standard deviation of MKL-DR) made in the case of caltech101 hold true for oxford flowers17 dataset also. Figures 3 and 4 respectively show the kernel weights for MKL-RT-KFDA and MKL-DR-KFDA approaches for the third split.

³ <http://www.robots.ox.ac.uk/~vgg/data/flowers/17/index.html>

Table 5: Average recognition rates for Oxford flowers17 dataset.

NN		KFDA + NN	
AK	71.9 \pm 1.5	AK-RT	86.3 \pm 1.5
PK	72.7 \pm 2.1	PK-RT	86.7 \pm 1.4
BIK	62.1 \pm 1.9	BIK-RT	73.8 \pm 2.0
		MKL-RT	87.3 \pm 1.5
		MKL-DR	84.8 \pm 1.1
SVM		KFDA + SVM	
AK	84.9 \pm 1.9	AK-RT	86.1 \pm 1.7
PK	85.5 \pm 1.2	PK-RT	86.8 \pm 1.6
BIK	70.6 \pm 1.6	BIK-RT	74.1 \pm 2.1
MKL	85.2 \pm 1.5	MKL-RT	87.3 \pm 1.7
		MKL-DR	84.8 \pm 1.1

Fig. 3: Weights of proposed convex MKL-RT-KFDA for Oxford flowers17.

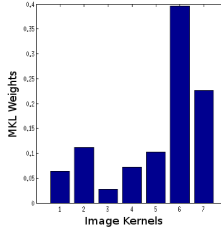
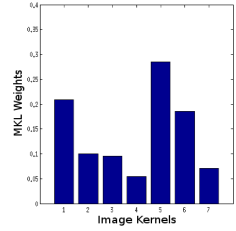


Fig. 4: Weights of non-convex MKL-DR-KFDA for Oxford flowers17.



5.3 KCCA for Cross-modal Retrieval

In these experiments, we used KCCA to map the data from two different modalities (image and text) to a common latent space, and used the cosine distance in the latent space for cross-modal retrieval. We used Wikipedia articles [30] dataset for these experiments. We measure the retrieval performance using mean average precision (MAP) [30].

Wikipedia articles [30] is a dataset of image-text pairs designed for cross-modal retrieval applications. It consists of 2173 training image-text pairs and 693 test image-text pairs which are grouped into 10 broad categories like art, history, etc. For text, we used a linear kernel generated from the 10-dimensional latent Dirichlet allocation model-based features provided by [30].⁴ For images, we extracted various features and constructed 21 different kernels as described below:

PHOG shape descriptor [4]: The descriptor is a histogram of oriented (Shp_{360}) or unoriented (Shp_{180}) gradients computed on the output of a Canny edge detector. The Shp_{360} histogram consists of 40 bins and the Shp_{180} histogram consists of 20 bins. We generated 4 kernels corresponding to different levels of spatial pyramid [22] from both Shp_{360} and Shp_{180} . Each kernel is an RBF kernel based on the χ^2 distance between histograms.

SIFT appearance descriptor: We computed the grayscale SIFT [24] descriptors on a regular grid on the image with a spacing of 2 pixels and for four different sizes $r = 4, 6, 8, 10$. We followed two different approaches, namely BOW model with 1000 codewords and second-order pooling [6], to obtain region descriptors from the SIFT descriptors. In each case, we generated 3 kernels corresponding to different levels of spatial pyramid. In the case of BOW model, each kernel is an RBF kernel based on the χ^2 distance between histograms. In the case of second-order pooling, each kernel is an RBF kernel based on the log-Euclidean distance [1] between covariance matrices.

LBP texture features [27]: We used the histograms of uniform rotation invariant $LBP_{8,1}$ features and generated 3 kernels corresponding to different levels of spatial pyramid. Each kernel is an RBF kernel based on the χ^2 distance between histograms.

⁴ <http://www.svcl.ucsd.edu/projects/crossmodal/>

Table 6: MAP scores for Wikipedia dataset.

Method	Query		Average
	Text	Image	
AK-RT-KCCA	27.74	31.43	29.58
PK-RT-KCCA	27.86	31.49	29.68
BIK-RT-KCCA	26.45	30.4	28.43
MKL-RT-KCCA	28.14	32.73	30.43
MKL-DR-KCCA	27.44	30.74	29.09

Fig. 5: Weights of proposed convex MKL-RT-KCCA for Wikipedia.

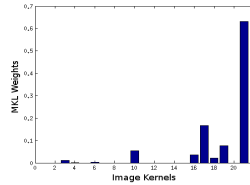
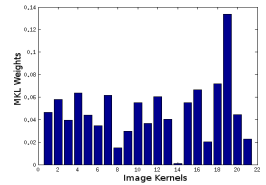


Fig. 6: Weights of non-convex MKL-DR-KCCA for Wikipedia.



Region covariance features [34]: We used the covariance of simple per-pixel features described in [34]. We generated 3 kernels corresponding to different levels of spatial pyramid. Each kernel is an RBF kernel based on the Log-Euclidean distance [1] between covariance matrices.

GIST image descriptor [28]: We generated an RBF kernel using the 512-dimensional GIST descriptor that records the pooled steerable filter responses within a grid of spatial cells across the image.

The number of KCCA dimensions was chosen to be 9 and the regularization parameter σ was chosen using cross-validation. Table 6 shows the MAP scores of various KCCA approaches on this dataset for text and image queries. We can clearly see that the proposed MKL-RT approach gives the best retrieval performance. Similar to KFDA experiments, the MKL-DR approach performs poorly compared to AK-RT, PK-RT and MKL-RT. Figures 5 and 6 respectively show the kernel weights for MKL-RT-KCCA and MKL-DR-KCCA approaches. For this dataset, the proposed convex MKL-RT-KCCA approach selected 9 kernels out of 21, whereas the non-convex MKL-DR-KCCA approach ended up selecting all the kernels. This clearly shows that the proposed approach can be successfully used for feature selection in cross-modal retrieval applications.

5.4 LKCCA for Cross-modal Retrieval

In these experiments, we used LKCCA to map the data from two different modalities (image and text) to a common latent space, and used the cosine distance in the latent space for cross-modal retrieval. We used Wikipedia articles [30] and PascalVOC 2007 [16] datasets for these experiments. The number of LKCCA dimensions was chosen to be $P - 1$, where P is the number of classes. We use the mean average precision to measure the retrieval performance.

For the Wikipedia dataset, we used the same image and text kernels that were used in KCCA experiments. Table 7 shows the MAP scores of various LKCCA approaches on this dataset for text and image queries. We can clearly see that the proposed MKL-RT approach gives the best retrieval performance. Similar to KFDA and KCCA experiments, the MKL-DR approach performs poorly compared to AK-RT, PK-RT and MKL-RT. Figures 7 and 8 respectively show the kernel weights for MKL-RT-LKCCA and MKL-DR-LKCCA approaches. For this dataset, the proposed convex MKL-RT-LKCCA approach selected 12 kernels out of 21, whereas the non-convex MKL-DR-LKCCA approach selected 18 kernels.

Table 7: MAP scores for Wikipedia dataset.

Method	Query		Average
	Text	Image	
AK-RT-LKCCA	27.28	31.02	29.15
PK-RT-LKCCA	27.16	30.75	28.96
BIK-RT-LKCCA	26.43	30.45	28.44
MKL-RT-LKCCA	27.92	32.09	30.01
MKL-DR-LKCCA	26.73	29.1	27.92

Fig. 7: Weights of proposed convex MKL-RT-LKCCA for Wikipedia.

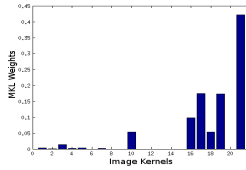


Fig. 8: Weights of non-convex MKL-DR-LKCCA for Wikipedia.

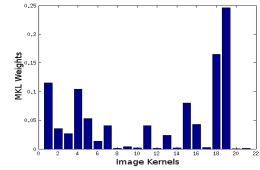


Table 8: MAP scores for PascalVOC dataset

Method	Query		Average
	Text	Image	
AK-RT-LKCCA	53.1	51.41	52.26
PK-RT-LKCCA	52.05	50.87	51.46
BIK-RT-LKCCA	51.81	51.39	51.6
MKL-RT-LKCCA	53.56	52.92	53.24
MKL-DR-LKCCA	55.05	46.49	50.77

Fig. 9: Weights of proposed convex MKL-RT-LKCCA for PascalVOC.

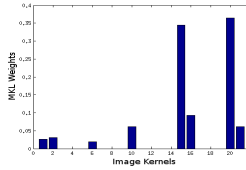
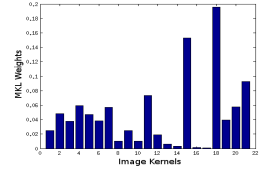


Fig. 10: Weights of non-convex MKL-DR-LKCCA for PascalVOC.



PascalVOC 2007 [16] dataset consists of 5011 training image-text pairs and 4952 test image-text pairs corresponding to 20 different object categories. Since some of the images are multi-labeled, following [32,38], we selected the images with only one object. This gave us 2799 training image-text pairs and 2820 test image-text pairs. For text, we used a linear kernel generated from the *absolute* and *relative* tag rank features provided by [16].⁵ For images, we extracted various different features (same as those used for the Wikipedia dataset) and constructed 21 different kernels.

Table 8 shows the MAP scores of various LKCCA approaches on this dataset for text and image queries. For image query, the proposed MKL-RT approach gives the best retrieval performance and the MKL-DR approach performs very poorly. For text query, the MKL-DR approach gives the best performance and the proposed MKL-RT approach is the second best. Considering the average retrieval performance, the proposed MKL-RT approach is the best. Figures 9 and 10 respectively show the kernel weights for MKL-RT-LKCCA and MKL-DR-LKCCA approaches. For this dataset, the proposed convex MKL-RT-LKCCA approach selected 8 kernels out of 21, whereas the non-convex MKL-DR-LKCCA approach selected 20 kernels.

For PascalVOC dataset, we also performed experiments using KCCA. The results produced by all the KCCA methods were much lower than the corresponding results of LKCCA methods. So, in the interest of space, we are not presenting those results.

6 Conclusion and Future Work

In this paper, we showed that MKL can be formulated as a convex optimization problem for a large class of ratio-trace problems that includes many popular algorithms like LDA, SDA, SILDA, LDE, MFA, LPP, NPE, CCA and Orthonormal PLS-SB. We also

⁵ <http://vision.cs.utexas.edu/projects/tag/bmvc10.html>

provided an optimization procedure that is guaranteed to converge to the global optimum of the proposed convex optimization problem. We performed experiments using three different instances of the ratio-trace problem and demonstrated that the proposed MKL-RT approach can be successfully used to select features for discriminative dimensionality reduction and cross-modal retrieval. We also showed that the proposed convex MKL-RT approach performs better than the non-convex MKL-DR approach of [23].

In the near future, we plan to test our approach on various other instances of the ratio-trace problem. Similar to the lines of ℓ_p -MKL-SVM and ℓ_p -MKL-LDA, we also plan to extend this work to ℓ_p -MKL-RT.

References

1. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Mat. Anal. App.* 29(1), 328–347 (2007)
2. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. *Neural Computation* 12(10), 2385–2404 (2000)
3. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *PAMI* 19(7), 711–720 (1997)
4. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: *CIVR* (2007)
5. Cai, D., He, X., Han, J.: Semi-supervised discriminant analysis. In: *ICCV* (2007)
6. Carreira, J., Caseiro, R., Batista, J., Sminchisescu, C.: Semantic segmentation with second-order pooling. In: *ECCV* (2012)
7. Chen, H.T., Chang, H.W., Liu, T.L.: Local discriminant embedding and its variants. In: *CVPR* (2005)
8. Etemad, K., Chellappa, R.: Discriminant analysis for recognition of human face images. *Jl. Optical Society of America* 14, 1724–1733 (1997)
9. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: *CVPR Workshops* (2004)
10. Fukunaga, K.: *Introduction to statistical pattern recognition*. Academic Press, second edn. (1991)
11. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: *ICCV* (2009)
12. Gönen, M., n, E.A.: Multiple kernel learning algorithms. *JMLR* 12, 2211–2268 (2011)
13. Hardoon, D.R., Szedmak, S., S.-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. *Neural Computation* 16(12), 2639–2664 (2004)
14. He, X., Cai, D., Yan, S., Zhang, H.J.: Neighborhood preserving embedding. In: *ICCV* (2005)
15. He, X., Niyogi, P.: Locality preserving projections. In: *NIPS* (2003)
16. Hwang, S.J., Grauman, K.: Reading between the lines: Object localization using implicit cues from image tags. *PAMI* 34(6), 1145–1158 (2012)
17. Kan, M., Shan, S., Xu, D., Chen, X.: Side-information based linear discriminant analysis for face recognition. In: *BMVC* (2011)
18. Kim, S.J., Magnani, A., Boyd, S.: Optimal kernel selection in kernel fisher discriminant analysis. In: *ICML* (2006)
19. Kim, T.K., Wong, S.F., Cipolla, R.: Tensor canonical correlation analysis for action classification. In: *CVPR* (2007)

20. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: lp-norm multiple kernel learning. *JMLR* 12, 953–997 (2011)
21. Lanckriet, G.R.G., Cristianini, N., Bartlett, P., Ghaoui, L.E., Jordan, M.I.: Learning the kernel matrix with semidefinite programming. *JMLR* 5, 27–72 (2004)
22. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *CVPR* (2006)
23. Lin, Y.Y., Liu, T.L., Fuh, C.S.: Multiple kernel learning for dimensionality reduction. *PAMI* 33(6), 1147–1160 (2011)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *IJCV* 60(2), 91–110 (2004)
25. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: *CVPR* (2006)
26. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: *ICVGIP* (2008)
27. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *PAMI* 24(7), 971–987 (2002)
28. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV* 42(3), 145–175 (2001)
29. Rakotomamonjy, A., Bach, F.R., Canu, S., Grandvalet, Y.: SimpleMKL. *JMLR* 9, 2491–2521 (2008)
30. Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: *ACM Multimedia* (2010)
31. Rosipal, R., Krämer, N.: Overview and recent advances in partial least squares. *Lecture Notes in Computer Science* pp. 34–51 (2006)
32. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: *CVPR* (2012)
33. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. *JMLR* 7, 1531–1565 (2006)
34. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: *ECCV* (2006)
35. Vedaldi, A., Gulshan, V., Varma, M., Zisserman, A.: Multiple kernels for object detection. In: *ICCV* (2009)
36. Vemulapalli, R., Pillai, J.K., Chellappa, R.: Kernel learning for extrinsic classification of manifold features. In: *CVPR* (2013)
37. Wang, H., Yan, S., Xu, D., Tang, X., Huang, T.S.: Trace ratio vs. ratio trace for dimensionality reduction. In: *CVPR* (2007)
38. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: *ICCV* (2013)
39. Yan, F., Kittler, J., Mikolajczyk, K., Tahir, A.: Non-sparse multiple kernel fisher discriminant analysis. *JMLR* 13, 607–642 (2012)
40. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: A general framework for dimensionality reduction. *PAMI* 29(1), 40–51 (2007)
41. Ye, J., Ji, S., Chen, J.: Multi-class discriminant kernel learning via convex programming. *JMLR* 9, 719–758 (2008)
42. Yeh, Y.R., Lin, T.C., Chung, Y.Y., Wang, Y.C.F.: A novel multiple kernel learning framework for heterogeneous feature fusion and variable selection. *IEEE Transactions on Multimedia* 14(3), 563–574 (2012)